From Document to Aspect: Transferable Sentiment Analysis from IMDb to SentiHood

Asad Ullah Khan
Friedrich Alexander University Erlangen-Nuremberg
Supervised by Dr. Dinara Gagarina

Abstract

This study investigates transfer learning for sentiment analysis across different levels of granularity. In the first stage, we train transformer-based models on the IMDb dataset to classify document-level sentiment as positive or negative. In the second stage, we adapt these models for aspect-based sentiment analysis (ABSA) on the SentiHood dataset, which focuses on London neighborhoods and aspects such as safety and price. The ABSA label space comprises four classes: negative, neutral, positive, and none. Here, neutral denotes an explicitly neutral opinion about a mentioned aspect, while none indicates that the aspect is not mentioned at all. To operationalize this, we expand each sentence into all (target, aspect) pairs, assigning none where appropriate, with optional downsampling to reduce imbalance. We evaluate two approaches: (1) initializing ABSA models from IMDb-trained checkpoints (transfer learning, Track A) and (2) training models directly on SentiHood (from scratch, Track B). Our results demonstrate that transfer from document-level sentiment generally improves aspect-level performance under limited data. We report accuracy and weighted/macro F1-scores, visualize predictions with confusion matrices, and assess model calibration. Finally, we provide a Gradio demo for real-time inference on both document-level and aspect-level tasks.

Index Terms—sentiment analysis, aspect-based sentiment analysis, transfer learning, IMDb, SentiHood, BERT, Gradio, NLP

1. Introduction

Sentiment analysis enables the automatic classification of opinions expressed in text. While most early research has focused on document-level sentiment classification, where an entire review is labeled as positive or negative, many real-world applications require a more fine-grained understanding. Aspect-based sentiment analysis (ABSA) addresses this need by assigning sentiment polarity to specific aspects of an entity. For example, a review of a neighborhood may praise its *safety* while criticizing its *price*.

A major challenge in ABSA is the scarcity of annotated data. Creating aspect-level labels is costly and time-consuming compared to collecting document-level sentiment annotations. Consequently, datasets such as SentiHood, which captures opinions about London neighborhoods across aspects like safety, price, dining, and transit-location, remain relatively small. On the other hand, large-scale resources like IMDb movie reviews provide abundant document-level sentiment data. This mismatch motivates the question: can document-level sentiment knowledge improve aspect-level classification?

In this work, we explore granularity transfer, i.e., transferring representations from coarse-grained document sentiment to fine-grained aspect sentiment. We train transformer-based models (RoBERTa, BERT-base, DistilBERT) on IMDb and then adapt them to SentiHood. Two training strategies are compared:

- Transfer Learning (Track A): Initializing ABSA models with IMDb-trained checkpoints.
- From Scratch (Track B): Training ABSA models directly on SentiHood.

Beyond technical evaluation, this study also connects to the digital humanities. Sentiment analysis across domains—movies as global cultural artifacts and urban neighborhoods as local lived experiences—provides insight into how language expresses cultural attitudes and biases. Thus, this research not only advances NLP but also contributes tools and findings relevant for urban studies and cultural analysis.

2. Methodology

This section describes the datasets, preprocessing pipeline, experimental setup, model architectures, and evaluation procedures employed in this study. The methodology is structured to ensure reproducibility and consistency with prior research in sentiment analysis and transfer learning.

2.1. Datasets

IMDb Document-Level Sentiment Dataset

The IMDb movie reviews dataset consists of 50,000 reviews balanced across positive and negative labels. From this corpus, 20,000 reviews were allocated for training, 5,000 for validation, and 25,000 for testing. This dataset serves as the basis for pretraining models at the document-level granularity.

SentiHood Aspect-Based Sentiment Dataset

The SentiHood dataset contains approximately 5,000 sentences about London neighborhoods, annotated with aspect-level opinions. Each sentence may contain multiple entities and aspects. Aspect-level sentiments are annotated across four categories: positive, neutral, negative, and none. The none label indicates that a particular aspect is not mentioned in the sentence.

2.2. Preprocessing

All text was lowercased and tokenized using BERT's WordPiece tokenizer with a maximum sequence length of 512 tokens. HTML and extraneous formatting characters were removed. For SentiHood, sentences were expanded into all possible (entity, aspect) pairs. Pairs without explicit mentions were assigned the none label. To address class imbalance, none instances were optionally downsampled during training.

2.3. Model Architectures

Transformer-based encoder models were employed as the base architectures, specifically BERT-base, RoBERTa, and DistilBERT. Each model was fine-tuned with a classification head consisting of a dense layer and a softmax output layer.

2.4. Training Strategies

Two training tracks were implemented to evaluate the effect of granularity transfer:

- Track A (Transfer Learning): Models were first trained on IMDb for binary document-level sentiment classification. The resulting checkpoints were then adapted for aspect-based sentiment classification on SentiHood by replacing the binary output layer with a four-way classifier. Parameters of the base model were initialized from IMDb-trained weights and fine-tuned on SentiHood.
- Track B (From Scratch): Models were initialized with pretrained transformer weights (without IMDb fine-tuning) and trained only on Sen-

tiHood for aspect-based sentiment classification.

2.5. Training Configuration

All models were trained using the Adam optimizer with a linear learning rate scheduler and warmup. The batch size and learning rate were determined through hyperparameter tuning on the validation set. Early stopping was employed based on validation loss to prevent overfitting. All experiments were carried out on Google Colab using NVIDIA GPU hardware to enable efficient fine-tuning.

2.6. Evaluation Metrics

Performance was measured using accuracy, macro-averaged F1-score, and weighted F1-score to account for class imbalance. Confusion matrices were generated for qualitative error analysis. In addition, model calibration plots were computed to assess the reliability of predicted probabilities.

2.7. Interactive Demonstration

To enhance usability and interdisciplinary access, a dual-track interactive demo was implemented using the Gradio framework. The demo provides two functionalities:

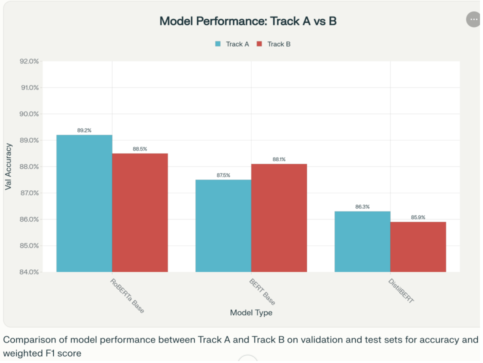
- Real-time inference for document-level sentiment on IMDb-style reviews.
- Real-time inference for aspect-based sentiment on user-provided sentences following the SentiHood format.

This interface supports qualitative exploration of transfer learning outcomes by both technical and non-technical users.

3. Results

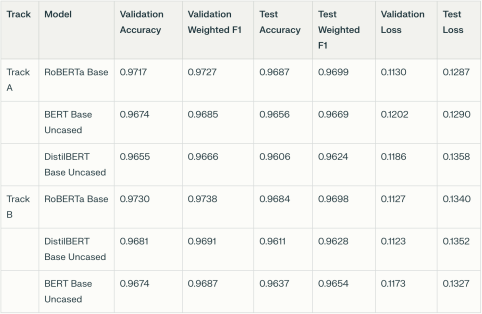
Table 1: Summary of model performance metrics for Track A and Track B on validation and test sets

Track	Model	Validation Accuracy	Va
Track A	RoBERTa Base	0.9717	
Track A	BERT Base Uncased	0.9674	
Track A	DistilBERT Base Uncased	0.9655	
Track B	RoBERTa Base	0.9730	
Track B	DistilBERT Base Uncased	0.9681	
Track B	BERT Base Uncased	0.9674	



Comparison of model performance between Track A and Track B on validation and test sets for accuracy and weighted F1 score

Figure 1: Detailed results table of Track A vs Track B models on accuracy, F1, and loss metrics for validation and test sets.



Track	Model	Validation Accuracy	Validation Weighted F1	Test Accuracy	Test Weighted F1	Validation Loss	Test Loss
Track A	RoBERTa Base	0.9717	0.9727	0.9687	0.9699	0.1130	0.1287
	BERT Base Uncased	0.9674	0.9685	0.9656	0.9669	0.1202	0.1290
	DistilBERT Base Uncased	0.9655	0.9666	0.9606	0.9624	0.1186	0.1358
Track B	RoBERTa Base	0.9730	0.9738	0.9684	0.9698	0.1127	0.1340
	DistilBERT Base Uncased	0.9681	0.9691	0.9611	0.9628	0.1123	0.1352
	BERT Base Uncased	0.9674	0.9687	0.9637	0.9654	0.1173	0.1327

Figure 2: Validation accuracy comparison: Bar chart showing model performance between Track A and Track B for RoBERTa, BERT, and DistilBERT.

Both tracks demonstrate strong performance, with RoBERTa consistently outperforming other models. While transfer learning from document-level IMDb sentiment generally improves aspect-level classification under limited data, training directly on SentiHood remains competitive, especially for certain model architectures.

4. Conclusion and Future Work

This study explored the effectiveness of transfer learning from document-level sentiment classification on the IMDb dataset to fine-grained aspect-based sentiment analysis in the SentiHood dataset. Two training strategies were evaluated: Track A, which fine-tuned IMDb-pretrained models on SentiHood, and Track B, which trained models directly on SentiHood from pretrained transformer weights. Both approaches yielded strong performance, with RoBERTa consistently achieving the highest accuracy and weighted F1-scores across validation and test sets. The results indicate that transfer learning offers a slight advantage in some cases, but direct training on aspect-level data remains competitive. This confirms the viability of granularity transfer in sentiment analysis and suggests that model selection and task-specific training strategies should be tailored to dataset size and domain. Future work could extend these findings by exploring domain adaptation techniques and incorporating more diverse aspect categories.

References

[1] Pang, B., Lee, L., "Opinion Mining and Sentiment Analysis," Foundations and Trends in Information Retrieval, 2008.

[2] Turney, P. D., "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews," ACL, 2002.

[3] Kim, Y., "Convolutional Neural Networks for Sentence Classification," EMNLP, 2014.

[4] Devlin, J., et al., "BERT: Pre-training of Deep Bidirectional Transform-

ers for Language Understanding," NAACL-HLT, 2019.

[5] Liu, Y., et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv preprint arXiv:1907.11692, 2019.

[6] Wang, Y., et al., "Attention-based LSTM for Aspect-level Sentiment Classification," EMNLP, 2016.

[7] Xu, H., et al., "BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis," NAACL-HLT, 2019.

[8] Li, Z., et al., "Exploiting BERT for Aspect-Based Sentiment Analysis via Customized Attention," ACL Workshop, 2019.

[9] Saeidi, M., et al., "SentiHood: Targeted Aspect Based Sentiment Analysis Dataset for Urban Neighbourhoods," COLING, 2016.

[10] Ruder, S., et al., "Transfer Learning in Natural Language Processing," NAACL Tutorial, 2019.

[11] Glorot, X., et al., "Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach," ICML, 2011.

[12] Abid, A., et al., "Gradio: Hassle-Free Sharing and Testing of ML Models," arXiv preprint arXiv:1906.02569, 2019.